
**Electronic Journal of Applied Statistical Analysis**
**EJASA, Electron. J. App. Stat. Anal.**
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v8n2p136

# Adaptive inference for multi-stage unbalanced exponential survey data based on a simulation from an intercept-only mode

By Al-Zou'bi

Published: 14 October 2015

*Electronic Journal of Applied Statistical Analysis*
Vol. 08, Issue 02, October 2015, 136-153
DOI: 10.1285/i20705948v8n2p136

# Adaptive inference for multi-stage unbalanced exponential survey data based on a simulation from an intercept-only mode

Loai Mahmoud Awad Al-Zou'bi*

*Al al-Bayt University, Department of Mathematics*
*Mafraq (25113), Jordan*

Published: 14 October 2015

Two-stage sampling usually leads to higher variances for estimators of means and regression coefficients, because of intra-cluster homogeneity. One way of allowing for clustering in fitting a linear regression model is to use a linear mixed model with two levels. If the estimated intra-cluster correlation is close to zero, it may be acceptable to ignore clustering and use a single level model. In this paper, an adaptive strategy is evaluated for estimating the variances of estimated regression coefficients. The strategy is based on testing the null hypothesis that random effect variance component is zero. If this hypothesis is accepted the estimated variances of estimated regression coefficients are extracted from the one-level linear model. Otherwise, the estimated variance is based on the linear mixed model, or, alternatively the Huber-White robust variance estimator is used. A simulation study is used to show that the adaptive approach provides reasonably correct inference in a simple case.

**keywords:** Adaptive estimation, variance components, cluster sampling, multi-level models, Huber-White variance estimator, exponential distribution, unbalanced data.

*Corresponding author: loai67@yahoo.com


ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

# 1 Introduction

Multilevel models are generalization of regression models. Let $y_{ij}$ be a dependent variable of interest, and $\mathbf{x}_{ij}$ a vector of covariates for unit $j$ in primary sampling unit (PSU) $i$. The two-level linear mixed model (LMM) Goldstein (2003) is given by

$$y_{ij} = \boldsymbol{\beta}'\mathbf{x}_{ij} + b_i + e_{ij}, \quad i = 1, 2, \ldots, c, \quad j = 1, 2, \ldots, m_i \tag{1}$$

where $c$ denotes the number of PSUs in the sample, $m_i$ denotes the number of observations selected in PSU $i$, $\boldsymbol{\beta}$ is the vector of unknown regression coefficients, $b_i$ is a PSU specific random effect with variance $\sigma_b^2$, we assume that $b_i$ has an exponential distribution with parameter $\mu$, $b_i \sim Exp(\mu)$, and $e_{ij}$ is assumed to be $N(0, \sigma_e^2)$.

A complication of two-stage sampling is that values of a variable of interest may tend to be more similar for units from the same PSU than for units from different PSUs. The intraclass correlation (ICC), $\rho$, is a measure of the association between the observations for members of the same PSU. It also describes the PSU homogeneity [Chapter 6]. If the intraclass correlation is non-zero, the clustered nature of the design should be reflected in the analysis procedure. One way of doing this is by fitting a multilevel model (MLM) Goldstein (2003)[Chapter 1].

In practice the intraclass correlation is often quite small. For example, if units within PSUs are no more homogenous than units over all PSUs, then the intraclass correlation is zero. On the other hand, if units from the same PSU have equal values then the intraclass correlation is 1. The intraclass correlation may take a negative value, but in practice it is generally positive. If each PSU in the population contains $M$ units, the smallest possible value of $\rho$ is $-1/(M-1)$. This occurs when the population is finite with high heterogeneity within PSUs, and zero variance between PSU means [p.260], show this for repeated probability sampling from a fixed finite population].

In this paper we will focus on modeling two-stage survey data. In the case of unequal number of observations in each PSU, $\rho$ is usually less than 0.1 when PSUs are geographic areas and final units are households in these areas (Verma et al., 1980). When PSUs are households and final units are people in households it is usually between 0 and 0.2 (Clark and Steel, 2002).

In ALZOUBI(2011), the methods were based on fitting a linear mixed model. Data were assumed to be normally distributed. In this article, the purpose is to see if these methods still work well if the assumption of normality is not justified. For this purpose, the same methods applied in ALZOUBI(2011) will be applied to data that are exponentially distributed rather than normal.

Exponential distribution is encountered as life-time distribution with constant hazard rate $\mu$. It is commonly employed in the formation of models of lifetime distributions, and stochastic process in general. If $X$ is an exponential random variable, the probability density function of $X$ is

$$f(x) = \begin{cases} \mu e^{-\mu x} & x \geq 0, \\ 0 & otherwise. \end{cases}$$

GOMRY noted that the mean and the variance of the exponential random variable $X$; respectively are

$$\begin{aligned} E(X) &= \mu; \\ Var(X) &= \mu^2. \end{aligned} \tag{2}$$

The exponential distribution is the only continuous memoryless random distribution.

## 2 Fitting the linear mixed model

### 2.1 The model

Let $\mathbf{X}$ be the $n \times p$ design matrix, which is assumed to be of rank $p$, and $\mathbf{Y} = (\mathbf{y}_1', \ldots, \mathbf{y}_c')'$ be the complete set of $n = \sum_{i=1}^{c} m_i$ observations in the $c$ groups, where $\mathbf{y}_i = (y_{i1}, \ldots, y_{im_i})'$ is the observed vector for the $i^{th}$ PSU. Model (1) can also be written as

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}), \tag{3}$$

where $\mathbf{V}$ is a block diagonal matrix, $\mathbf{V} = diag(\mathbf{V}_i, i = 1, \ldots, c)$, and

$$\mathbf{V}_i = \sigma_b^2 \mathbf{J}_{m_i} + \sigma_e^2 \mathbf{I}_{m_i}, \tag{4}$$

where $\mathbf{J}_{m_i}$ is an $m_i \times m_i$ matrix with all entries equal to 1, and $\mathbf{I}_{m_i}$ is the $m_i \times m_i$ identity matrix. $\boldsymbol{\beta}$ is the vector of unknown regression coefficients.

A simple special case of model (1) is the intercept-only model, this model includes just a grand mean parameter, it is defined by setting $x_{ij}$ to 1 for all $i, j$ as

$$y_{ij} = \beta + b_i + e_{ij}, \quad i = 1, 2, \ldots, c, \quad j = 1, 2, \ldots, m_i, \tag{5}$$

where $c$ denotes number of the sample PSUs, $m_i$ denotes the number of units selected in PSU $i$, $b_i \sim Exp(\mu)$ is a PSU specific random effect and $b_i s$ are independent and identically distributed ($iid$) with variance $\sigma_b^2$, and $e_{ij}$ is assumed to be $N(0, \sigma_e^2)$. The parameters $\sigma_b^2$ and $\sigma_e^2$ are the between- and within-PSUs variance components.

Observations for different units from the same PSU are correlated. It is assumed that $b_i$ is uncorrelated with $e_{ij}$, and that $b_i$ and $b_{i'}$ for $i \neq i'$ are uncorrelated. Therefore, as a consequence Rao (1997),

$$\begin{aligned} V(y_{ij}) &= V(b_i) + V(e_{ij}) = \sigma_b^2 + \sigma_e^2, \\ Cov(y_{ij}, y_{ij'}) &= V(b_i) = \sigma_b^2 \text{ for } j \neq j', \text{ and} \\ Cov(y_{ij}, y_{i'j}) &= 0 \text{ for } i \neq i'. \end{aligned} \tag{6}$$

## 2.2 Likelihood Theory Estimation of $var(\hat{\boldsymbol{\beta}})$

In this section we discuss the variances of the estimated regression coefficients and their estimators. The estimated variance of the restricted maximum likelihood (REML) of $\hat{\boldsymbol{\beta}}$ is given by

$$\begin{aligned} \widehat{var}(\hat{\boldsymbol{\beta}}) &= \left(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1} \\ &= \left(\sum_{i=1}^{c}\mathbf{x}_i'\hat{\mathbf{V}}_i^{-1}\mathbf{x}_i\right)^{-1}, \end{aligned} \tag{7}$$

where $\hat{\mathbf{V}}_i = \hat{\sigma}_b^2\mathbf{J}_{m_i} + \hat{\sigma}_e^2\mathbf{I}_{m_i}$. In the unbalanced data case, the intercept-only model given by (5) simplifies to

$$\widehat{var}(\hat{\beta}) = \Big\{\sum_{i=1}^{c}\frac{m_i}{\hat{\sigma}_e^2 + m_i\hat{\sigma}_b^2}\Big\}^{-1} = (\sum_{i=1}^{c}\hat{\lambda}_i)^{-1}, \tag{8}$$

A confidence interval for $\beta$ could be constructed using the Equation

$$(1-\alpha)100\%CI = \hat{\beta} \pm t_{(df,1-\frac{\alpha}{2})}\sqrt{\widehat{var}(\hat{\beta})}. \tag{9}$$

However, it is not clear how the degrees of freedom in (9) should be defined for mixed models. Faes et al. (2009) suggested using the effective sample size ($\nu$) as degrees of freedom for mixed models, with $\hat{\nu} = \frac{n}{\widehat{deff}(\hat{\beta})}$. The effective sample size is the ratio of the sample size to the design effect of $\hat{\beta}$. Other approaches have been suggested, see for example SAT and KR97. The method of Faes et al. (2009) has the advantage that it extends naturally to non-Gaussian model, unlike the other approaches.

## 2.3 Huber-White Estimator of $var(\hat{\boldsymbol{\beta}})$

Liang and Zeger (1986) suggested the generalized estimation equation (GEE) approach as an alternative to the ML and REML approaches for modeling longitudinal and cross-sectional data. The GEE approach to linear modeling of clustered data can use either ordinary least squares ($OLS$) or generalized least squares ($GLS$).

The $OLS$ estimator for $\boldsymbol{\beta}$ is defined by

$$\hat{\boldsymbol{\beta}}_{ols} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{10}$$

When the observations from different PSUs are uncorrelated but the same PSU observations are correlated with common intraclass correlation $\rho$, the estimator $\hat{\boldsymbol{\beta}}_{ols}$ is unbiased Scott and Holt (1982) with variance equal to

$$var(\hat{\boldsymbol{\beta}}_{ols}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \tag{11}$$

In general, $\mathbf{V}$ is not known and it can be estimated by $\hat{\mathbf{V}}$, therefore the estimated variance for $\hat{\boldsymbol{\beta}}_{ols}$ is defined by

$$\widehat{var}(\hat{\boldsymbol{\beta}}_{ols}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \tag{12}$$

The estimator $\widehat{var}(\hat{\boldsymbol{\beta}})$ in (7) is approximately unbiased provided that the variance model (4) is correct. Otherwise, $\widehat{var}(\hat{\boldsymbol{\beta}})$ will be biased and inference will be incorrect. An alternative to ML or REML estimates of $var(\hat{\boldsymbol{\beta}})$ is the robust variance estimate approach described by Liang and Zeger (1986), in the context of modeling longitudinal data using generalized estimating equations. This approach can be applied to the analysis of data collected using PSUs, where observations within PSUs might be correlated and the observations in different PSUs are independent.

This approach can be referred to as robust or Huber-White variance estimation (Huber, 1967; White, 1982). It will be used as an alternative approach to estimating $var(\hat{\boldsymbol{\beta}})$ in this paper. The method yields asymptotically consistent covariance matrix estimates even if the variances and covariances assumed in model (1) are incorrect. It is still necessary to assume that observations from different PSUs are independent.

In Equation (7), the variance of $\hat{\boldsymbol{\beta}}$ was estimated by substituting REML estimates of $\sigma_b^2$ and $\sigma_e^2$ into $\mathbf{V}_i$. An alternative estimator of $\mathbf{V}_i$ is $\hat{\mathbf{V}}_i^{Hub} = \hat{\mathbf{e}}_i\hat{\mathbf{e}}_i'$, where $\hat{\mathbf{e}}_i = \mathbf{y}_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}$. $\hat{\mathbf{V}}_i^{Hub}$ is approximately unbiased for $\mathbf{V}_i$ even if (4) does not apply.

$$\begin{aligned} E\left(\hat{\mathbf{V}}_i^{Hub}\right) &= E(\hat{\mathbf{e}}_i\hat{\mathbf{e}}_i') \\ &\approx E[(\mathbf{y}_i - \mathbf{x}_i'\boldsymbol{\beta})(\mathbf{y}_i - \mathbf{x}_i'\boldsymbol{\beta})'] \\ &= \mathbf{V}_i. \end{aligned} \tag{13}$$

Note that

$$\begin{aligned} var(\hat{\boldsymbol{\beta}}) &= var\big(\big(\textstyle\sum_{i=1}^{c}\mathbf{x}_i'\hat{\mathbf{V}}_i^{-1}\mathbf{x}_i\big)^{-1}\big(\textstyle\sum_{i=1}^{c}\mathbf{x}_i'\hat{\mathbf{V}}_i^{-1}\mathbf{y}_i\big)\big) \\ &\approx \big(\textstyle\sum_{i=1}^{c}\mathbf{x}_i'\hat{\mathbf{V}}_i^{-1}\mathbf{x}_i\big)^{-1}\big(\textstyle\sum_{i=1}^{c}\mathbf{x}_i'\hat{\mathbf{V}}_i^{-1}\mathbf{V}_i\hat{\mathbf{V}}_i^{-1}\mathbf{x}_i\big) \\ &\quad \big(\textstyle\sum_{i=1}^{c}\mathbf{x}_i'\hat{\mathbf{V}}_i^{-1}\mathbf{x}_i\big)^{-1}. \end{aligned} \tag{14}$$

One way to construct a robust estimator of $var(\hat{\boldsymbol{\beta}})$ is to substitute the robust estimator $\hat{\mathbf{V}}_i^{Hub}$ in (14) as follows (Liang and Zeger, 1986),

$$\begin{aligned} \widehat{var}_{Hub}(\hat{\boldsymbol{\beta}}) &= \Big(\textstyle\sum_{i=1}^{c}\mathbf{x}_i'\hat{\mathbf{V}}_i^{-1}\mathbf{x}_i\Big)^{-1}\Big(\textstyle\sum_{i=1}^{c}\mathbf{x}_i'\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{V}}_i^{Hub}\hat{\mathbf{V}}_i^{-1}\mathbf{x}_i\Big) \\ &\quad \Big(\textstyle\sum_{i=1}^{c}\mathbf{x}_i'\hat{\mathbf{V}}_i^{-1}\mathbf{x}_i\Big)^{-1}. \end{aligned} \tag{15}$$

When there is only an intercept in the model ($\mathbf{x}_{ij}$=1), (15) becomes

$$\widehat{var}_{Hub}(\hat{\beta}) = \frac{\sum_{i=1}^{c}\hat{\lambda}_i^2(\bar{y}_{i.} - \hat{\beta})^2}{(\sum_{i=1}^{c}\hat{\lambda}_i)^2}. \tag{16}$$

Exact confidence intervals can then be calculated with degrees of freedom equal to $c$-1 MAC85.

### 2.4 Restricted Likelihood Ratio Test (RLRT)

A better option is to use REML estimators to derive the likelihood ratio test (LRT) statistic for testing $H_0 : \sigma_b^2 = 0$.

The problem of testing $H_0 : \sigma_b^2 = 0$ using the likelihood ratio test is discussed by Self and Liang (1987) using ML estimators for the variance components. Self and Liang (1987) allowed the true parameter values to be on the boundary of the parameter space, and showed that the large sample distribution of the likelihood ratio test is a mixture of $\chi^2$ distributions under nonstandard conditions assuming that response variables are *iid*. This assumption does not generally hold in linear mixed models, at least under the alternative hypothesis.

The restricted log-likelihood function is given by West et al. (2007)[p.28] as

$$\begin{aligned} \ell_R &= -\tfrac{1}{2}\big[(n-1)log(2\pi) + log|\mathbf{V}| + log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| \\ &\quad +\mathbf{Y}'\mathbf{V}^{-1}\{\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\}\mathbf{V}^{-1}\mathbf{Y}\big], \end{aligned} \tag{17}$$

where $\mathbf{V} = diag(\mathbf{V}_i)$ and $\mathbf{V}_i$ are given by (4). Maximizing (17) with respect to $\sigma_b^2$ and $\sigma_e^2$ gives the REML estimates of these parameters.

From (17), the restricted likelihood ratio test is given by

$$\begin{aligned} \Lambda &= -2\, log(RLRT) \\ &= 2 \overset{MAX}{H_A} \ell_R(\boldsymbol{\beta}, \sigma_b^2, \sigma_e^2) - 2 \overset{MAX}{H_0} \ell_R(\boldsymbol{\beta}, \sigma_b^2, \sigma_e^2). \end{aligned} \tag{18}$$

The large sample distribution of the likelihood ratio $\Lambda$ is a 50:50 mixture of $\chi^2$ distribution with 0 and 1 degrees of freedom as the parameter values fall on the boundary of the parameter space (Self and Liang, 1987).

## 3 Adaptive strategies

In this paper we consider two adaptive strategies. Both of them rely on the idea of testing the variance component $\sigma_b^2$ in model (1). If we reject $H_0 : \sigma_b^2 = 0$, we use the first adaptive strategy which is utilizing the LMM-REML estimators of $var(\hat{\beta})$ defined in Equation (7). On the other hand, if we accept $H_0$, then we assume that $\sigma_b^2 = 0$ and we fit the standard linear model with independent errors. This strategy is explained in Figure 1, where $\widehat{var}_{\mathrm{LM}}(\hat{\beta})$ is the estimator of $\mathrm{var}_{\mathrm{LM}}(\hat{\beta})$ using the LM strategy, $\widehat{var}_{\mathrm{LMM}}(\hat{\beta})$ is the estimator of $\mathrm{var}_{\mathrm{LMM}}(\hat{\beta})$ using the LMM strategy and $\widehat{var}_{\mathrm{ADM}}(\hat{\beta})$ is the adaptive estimator.

The second adaptive strategy, explained in figure 2, is identical, except that the robust Huber-White estimator $\widehat{var}_{Hub}(\hat{\beta})$ is used instead of $\widehat{var}_{LMM}(\hat{\beta})$ when $H_0$ is rejected.

The advantage of the adaptive strategy is that we use the simple linear model to derive variance estimators, unless there is strong evidence that $H_0 : \sigma_b^2 > 0$. This has benefit of simplifying the model and may also give tighter confidence intervals. However, it is not clear whether the adaptive approaches will give valid confidence intervals for $\beta$, because the confidence intervals assume non-adaptive procedures.

Figure 1: Flowchart explaining the adaptive procedure using the estimated variance extracted from the LMM

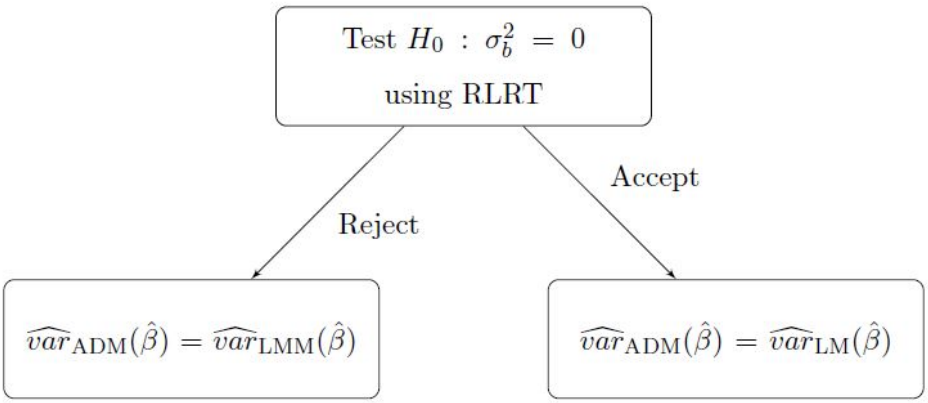
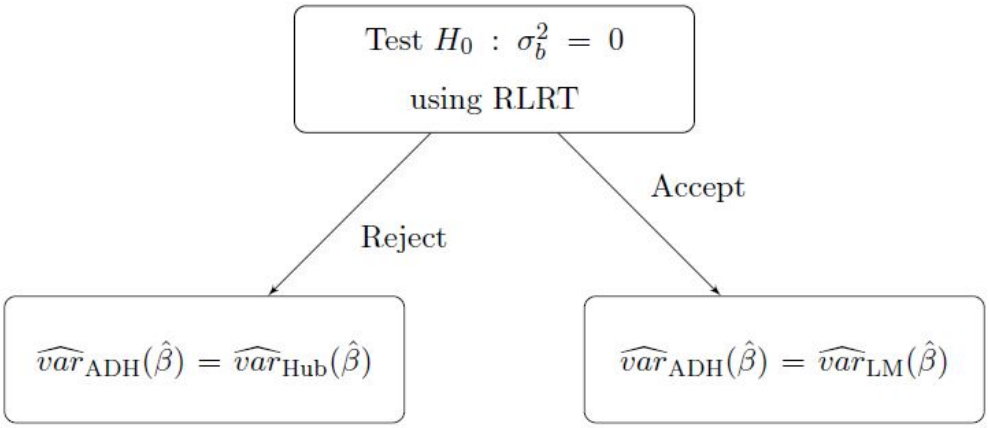



Figure 2: Flowchart explaining the adaptive procedure using Huber-White estimator



## 4 Simulation study

A simulation study was conducted to compare the adaptive and non-adaptive methods for estimating $var(\hat{\beta})$ using PSUs with unequal sample sizes. Data were generated from model (5), with different PSU sizes, $m_i$, assuming $b_i$ is exponentially distributed. The value of $\rho$ was varied over a range of values of 0, 0.025 and 0.1. The number of PSUs, $c$, was also varied over a range of values of 2, 5, 10, 25 and 50. $m_i$ generated randomly from uniform distribution. The average number of observations per PSU, $\bar{m}$ was varied to be 3, 10 and 25 to be consistent with the balanced data case. For this purpose three cases were used. In case 1, the number of observations was generated to be an integer between 2 and 4 with average equal to 3 observations per PSU. In case 2, this number varied from 5 to 15, with average equal to 10. Finally, in case 3, the average was 25, with $m_i$ varying between 15 and 35. In each case1000 samples were generated. The hypothesis $H_0 : \sigma_b^2 = 0$ was tested as described in Subsection 2.4 using the restricted likelihood ratio test defined by

$$\begin{aligned}
\Lambda &= -2\Big(\overset{MAX}{H_0}\ \ell_R - \overset{MAX}{H_A}\ \ell_R\Big) \\
&= ln(n) + (n-1)\, ln(MSE_0) + \frac{\sum_{i=1}^{c} m_i(\bar{y}_{i.} - \bar{y}_{..})^2}{MSE_0} \\
&\quad -(n-c)ln(MSE_A) - \sum_{i=1}^{c} ln(\hat{\eta}_i) - ln\Big(\sum_{i=1}^{c}(\hat{\lambda}_i)\Big) \\
&\quad -\sum_{i=1}^{c} \hat{\lambda}_i(\bar{y}_{i.} - \hat{\beta})^2, \qquad (19)
\end{aligned}$$

where $MSE_0 = \frac{1}{n-1}\sum_{i=1}^{c}\sum_{j=1}^{m_i}(y_{ij} - y_{..})^2$ is the mean squared error under the null hypothesis, $\sigma_b^2 = 0$ and $MSE_A = \hat{\sigma}_e^2$ is the mean squared error under the alternative hypothesis, $\sigma_b^2 > 0$ and $\eta_i = \sigma_e^2 + m_i\sigma_b^2$.

Tables 1 - 3 show the results for the unbalanced data case. They show the ratio of the mean estimated variance of $\hat{\beta}$, $E(\widehat{var}(\hat{\beta}))/var(\hat{\beta})$, using the four strategies of estimation (ADM, ADH, LMM and Huber) with $\rho = 0$, 0.025 and 0.1. In all tables we used $\beta = 0$ and significance level $\alpha = 0.1$ for testing $\sigma_b^2 = 0$. The tables show the non-coverage rates of 90% confidence intervals for $\beta$ and the average lengths of these confidence intervals. The proportion of samples where $H_0 : \sigma_b^2 = 0$ was rejected are also shown.

The variance estimators were generally approximately unbiased as most ratios were close to 1. There were some exceptions. The first was the LMM, ADM and ADH variance estimators, which tended to be biased when there were 10 or less sample PSUs with all average numbers of observations per PSU for $\rho$=0. For $\rho$=0.025, it tended to be biased when $c \leq 5$ with all $\bar{m}$ values and when there were 5 sample PSU with $\bar{m}$ was 3. For $\rho$=0.1, it tended to be biased when $c$ was 2 with all values of $\bar{m}$.

Non-coverage rates for $\beta$ were close to the nominal rate of 10% when $\rho$=0 for all methods except for the LMM method. The LMM non-coverage rates were a bit smaller than the nominal rate when $c$=2 with all average numbers of observations per PSU. The LMM non-coverage was good when there were 5 or more PSUs.

For $\rho \neq 0$, Huber non-coverage rate increases as the number of PSUs increases. For $\rho = 0$, Huber non-coverage rate was close to 10.

For $\rho$=0, the LMM, ADM and ADH non-coverage rates were close to the nominal rate for both values of $\sigma$, as in the normal data case (see ALZOUBI(2011), except when there were small number of sample PSUs (10 or less) with all average numbers of observations per PSU.

For $\rho$=0.025, the LMM and ADM non-coverage rates were much higher than the nominal rate for all sample PSUs with all average number of observations per PSU. Except when there $\bar{m} = 3$ with values of $c$ of 2 and 5. The ADH non-coverage rate was higher than the nominal rate when there were 2 sample PSUs with $\bar{m}$=25. In case of $\rho$=0.1, the LMM and ADM non-coverage rates were much higher than the nominal rate when $c \leq$10 and $\bar{m}$=10 or 25, and when $c$=50 with $\bar{m}$=3. The ADH non-coverage rate was about the same as the nominal rate in most cases except when $c$=5 with all values of $\bar{m}$ for $\rho$=0, when $c$=2 and 5 with $\bar{m}$=25 and 3, respectively when $\rho$=0.025 and when $c$=2 with $\bar{m}$=10 and 25, $c$=5 with $\bar{m}$=25 and when $c$=50 with $\bar{m}$=3 in case of $\rho$=0.1.

The ADM average lengths of confidence intervals for $\beta$ were similar to the LMM average lengths of confidence intervals for $\beta$ for $c \geq 5$ with all average numbers of observations per PSU for all values of $\rho$. When $c$=2, the ADM average lengths were about 6-12% shorter. The ADH average lengths of confidence intervals for $\beta$ were similar to the Huber average lengths of confidence intervals for $\beta$ for all sample PSUs with all values of $m$ and $\rho$ except when $c$=2, as the ADH average lengths were shorter than the Huber average lengths of order about 30-65%.

The proportions of samples where $H_0 : \sigma_b^2 = 0$ is rejected were generally much higher than 10% when $\rho$=0, and was very high 27% when $c$=5 and $\bar{m}$=3. This might be because the PSU sizes in the unbalanced design have a wide range, for example; for $\bar{m}$=25, the PSU sizes vary between 15 and 35. Or this might be because of the distribution of the RLRT. It was assumed that the distribution is a 50:50 mixture of $\chi_0^2$ and $\chi_1^2$ following Chernoff (1954) in the balanced and unbalanced designs. The 50:50 mixture of $\chi^2$ distribution of the likelihood ratio test might not perform well in the unbalanced designs because the response can not be divided into identically distributed sub-vectors as in Stram and Lee (1994). This approximation may not be a very good approximation in the unbalanced designs if the response is divided into small or moderate number of sub-vectors, even if the responses are independent SONJA.

For $\rho$=0, the average length of the 90% ADH confidence intervals for $\beta$ was shorter than the Huber (about 35%) when there were 2 sample PSUs with all values of $\bar{m}$. For $\rho$=0.025, the average length of the 90% ADH confidence intervals for $\beta$ was shorter than the Huber (30-35%) when there were 2 sample PSUs with all values of $\bar{m}$. For $\rho = 0.1$

the average length of the 90% ADH confidence intervals for $\beta$ was shorter than the Huber (20-30%) when there were 2 sample PSUs with all values of $\bar{m}$.

The proportions of samples where $H_0 : \sigma_b^2 = 0$ was rejected were higher than the nominal rate (10%). Possible reasons why these proportions are higher than 10% are discussed in Subsection 4.

Results on non-coverage and 90% confidence intervals length are also shown in graphical form in Figures 3 - 5. In the graphs we also include the LM strategy of estimation so that the effect of completely ignoring the clustered nature of the data can be examined.

Non-coverage rates for confidence intervals for $\beta$ were close to the nominal rate of 10% when $\rho = 0$ for all methods.

For $\rho \neq 0$, Huber non-coverage was close to 10% when there were 2 sample PSUs and higher, otherwise. Whereas, the LMM, ADM and ADH non-coverage rates were high, in general.

Table 1: Variance ratios, average length and non-coverage of the 90% confidence intervals for $\beta$, and power of testing $H_0 : \sigma_b^2 = 0$ using RLRT in the unbalanced data case with $\rho$=0.

| PSUs | Obs | $E(\widehat{var}(\hat{\beta}))/var(\hat{\beta})$ | | | | Non-Coverage of CI for $\beta$ (%) | | | | Pr(Rej $H_0$) (%) | Confidence Interval Length | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| c | $\bar{m}$ | ADM | ADH | LMM | Hub | ADM | ADH | LMM | Hub | Lrt | ADM | ADH | LMM | Hub |
| 2 | 3 | 1.346 | 1.346 | 1.441 | 1.055 | 8.8 | 8.8 | 7.8 | 10.7 | 18.1 | 2.26 | 2.906 | 2.398 | 4.353 |
| 2 | 10 | 1.447 | 1.447 | 1.542 | 1.064 | 7.7 | 7.6 | 6.7 | 9.4 | 14.9 | 0.933 | 1.467 | 0.971 | 2.359 |
| 2 | 25 | 1.493 | 1.494 | 1.577 | 1.056 | 7.7 | 7.6 | 7.2 | 10.2 | 12.4 | 0.546 | 0.864 | 0.563 | 1.41 |
| 5 | 3 | 1.224 | 1.224 | 1.240 | 1.034 | 8.2 | 7.8 | 7.8 | 10.5 | 27.8 | 1.00 | 1.047 | 1.008 | 1.048 |
| 5 | 10 | 1.160 | 1.160 | 1.164 | 0.923 | 8.2 | 7.9 | 8 | 10.1 | 21.9 | 0.52 | 0.557 | 0.521 | 0.560 |
| 5 | 25 | 1.214 | 1.215 | 1.218 | 0.995 | 7.9 | 7.3 | 7.9 | 9.3 | 26.0 | 0.329 | 0.359 | 0.33 | 0.366 |
| 10 | 3 | 1.133 | 1.133 | 1.134 | 0.999 | 7.8 | 7.7 | 7.8 | 9.9 | 26.4 | 0.656 | 0.668 | 0.657 | 0.651 |
| 10 | 10 | 1.132 | 1.132 | 1.132 | 0.994 | 8.8 | 8.6 | 8.8 | 10.7 | 21.0 | 0.352 | 0.361 | 0.352 | 0.357 |
| 10 | 25 | 1.157 | 1.157 | 1.157 | 0.994 | 9.0 | 8.8 | 9.0 | 11.5 | 20.1 | 0.221 | 0.227 | 0.221 | 0.223 |
| 25 | 3 | 1.093 | 1.093 | 1.093 | 1.03 | 9.3 | 9.2 | 9.3 | 9.9 | 14.9 | 0.394 | 0.396 | 0.395 | 0.391 |
| 25 | 10 | 1.054 | 1.054 | 1.054 | 0.996 | 9.8 | 9.7 | 9.8 | 10.6 | 13.6 | 0.216 | 0.217 | 0.216 | 0.216 |
| 25 | 25 | 1.063 | 1.063 | 1.063 | 0.985 | 9.2 | 9.2 | 9.2 | 11.3 | 13.8 | 0.136 | 0.137 | 0.136 | 0.135 |
| 50 | 3 | 1.121 | 1.121 | 1.121 | 1.099 | 7.9 | 7.9 | 7.9 | 7.7 | 5.8 | 0.272 | 0.272 | 0.272 | 0.272 |
| 50 | 10 | 0.964 | 0.964 | 0.964 | 0.93 | 11.4 | 11.4 | 11.4 | 11.5 | 6.2 | 0.149 | 0.149 | 0.149 | 0.148 |
| 50 | 25 | 1.108 | 1.108 | 1.108 | 1.061 | 8.0 | 8.0 | 8.0 | 9.4 | 12.5 | 0.095 | 0.095 | 0.095 | 0.094 |

Table 2: Variance ratios, average length and non-coverage of the 90% confidence intervals for $\beta$, and power of testing $H_0 : \sigma_b^2 = 0$ using RLRT in the unbalanced data case with $\rho$=0.025.

| PSUs | Obs | $E(\widehat{var}(\hat{\beta}))/var(\hat{\beta})$ | | | | Non-Coverage of CI for $\beta$ (%) | | | | Pr(Rej $H_0$) (%) | Confidence Interval Length | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| c | $\bar{m}$ | ADM | ADH | LMM | Hub | ADM | ADH | LMM | Hub | Lrt | ADM | ADH | LMM | Hub |
| 2 | 3 | 1.282 | 1.282 | 1.383 | 1.038 | 10.5 | 10.4 | 9.4 | 10.6 | 19.0 | 2.324 | 3.016 | 2.482 | 4.522 |
| 2 | 10 | 1.354 | 1.354 | 1.432 | 1.065 | 16.1 | 16.0 | 14.0 | 12.6 | 18.1 | 1.001 | 1.688 | 1.037 | 2.567 |
| 2 | 25 | 1.28 | 1.281 | 1.341 | 1.085 | 22.9 | 21.3 | 21.4 | 12.0 | 23.5 | 0.634 | 1.280 | 0.654 | 1.830 |
| 5 | 3 | 1.198 | 1.199 | 1.215 | 1.038 | 12.5 | 11.8 | 12.5 | 14.0 | 30.2 | 1.026 | 1.078 | 1.034 | 1.088 |
| 5 | 10 | 1.067 | 1.067 | 1.073 | 0.923 | 24.1 | 21.8 | 23.8 | 23.3 | 31.6 | 0.552 | 0.606 | 0.553 | 0.621 |
| 5 | 25 | 1.113 | 1.113 | 1.115 | 1.036 | 37.9 | 29.7 | 37.7 | 28.3 | 49.1 | 0.390 | 0.454 | 0.390 | 0.465 |
| 10 | 3 | 1.103 | 1.103 | 1.105 | 1.001 | 19.3 | 18.2 | 19.3 | 20.1 | 30.1 | 0.672 | 0.686 | 0.673 | 0.676 |
| 10 | 10 | 1.048 | 1.048 | 1.048 | 0.987 | 40.2 | 37.9 | 40.2 | 37.3 | 36.6 | 0.379 | 0.395 | 0.379 | 0.399 |
| 10 | 25 | 1.018 | 1.018 | 1.018 | 0.993 | 66.8 | 62.2 | 66.8 | 60.4 | 51.2 | 0.259 | 0.276 | 0.259 | 0.281 |
| 25 | 3 | 1.078 | 1.079 | 1.078 | 1.048 | 34.6 | 34.3 | 34.6 | 34.5 | 19.3 | 0.403 | 0.405 | 0.403 | 0.406 |
| 25 | 10 | 0.943 | 0.943 | 0.943 | 0.954 | 70.9 | 70.0 | 70.9 | 68.8 | 38.3 | 0.233 | 0.236 | 0.233 | 0.242 |
| 25 | 25 | 0.973 | 0.973 | 0.973 | 0.993 | 94.8 | 94.1 | 94.8 | 93.4 | 64.6 | 0.165 | 0.169 | 0.165 | 0.173 |
| 50 | 3 | 1.087 | 1.087 | 1.087 | 1.104 | 58.8 | 58.7 | 58.8 | 57.6 | 11.2 | 0.279 | 0.279 | 0.279 | 0.284 |
| 50 | 10 | 0.875 | 0.875 | 0.875 | 0.912 | 94.1 | 94.0 | 94.1 | 93.3 | 35.1 | 0.161 | 0.162 | 0.161 | 0.167 |
| 50 | 25 | 1.094 | 1.094 | 1.094 | 1.104 | 99.7 | 99.7 | 99.7 | 99.7 | 85.7 | 0.119 | 0.121 | 0.119 | 0.122 |

Table 3: Variance ratios, average length and non-coverage of the 90% confidence intervals for $\beta$, and power of testing $H_0 : \sigma_b^2 = 0$ using RLRT in the unbalanced data case with $\rho$=0.1.

| PSUs | Obs | $E(\widehat{var}(\hat{\beta}))/var(\hat{\beta})$ | | | | Non-Coverage of CI for $\beta$ (%) | | | | Pr(Rej $H_0$) (%) | Confidence Interval Length | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| c | $\bar{m}$ | ADM | ADH | LMM | Hub | ADM | ADH | LMM | Hub | Lrt | ADM | ADH | LMM | Hub |
| 2 | 3 | 1.195 | 1.195 | 1.283 | 1.017 | 16.1 | 16.0 | 13.9 | 10.4 | 22.5 | 2.560 | 3.432 | 2.723 | 4.939 |
| 2 | 10 | 1.161 | 1.161 | 1.212 | 1.020 | 27.9 | 25.7 | 25.9 | 13.8 | 26.0 | 1.222 | 2.304 | 1.260 | 3.169 |
| 2 | 25 | 1.120 | 1.120 | 1.152 | 1.069 | 37.2 | 30.3 | 35.2 | 12.4 | 39.5 | 0.900 | 2.190 | 0.923 | 2.682 |
| 5 | 3 | 1.134 | 1.134 | 1.150 | 1.033 | 25.3 | 23.2 | 24.9 | 23.5 | 38.1 | 1.109 | 1.178 | 1.118 | 1.199 |
| 5 | 10 | 1.003 | 1.003 | 1.007 | 0.953 | 48.4 | 40.7 | 48.0 | 39.4 | 53.3 | 0.669 | 0.768 | 0.671 | 0.788 |
| 5 | 25 | 1.015 | 1.015 | 1.015 | 1.003 | 64.8 | 49.8 | 64.7 | 48.1 | 73.8 | 0.555 | 0.672 | 0.555 | 0.682 |
| 10 | 3 | 1.050 | 1.050 | 1.051 | 0.998 | 44.8 | 42.7 | 44.8 | 42.9 | 43.5 | 0.730 | 0.751 | 0.731 | 0.751 |
| 10 | 10 | 0.995 | 0.995 | 0.995 | 0.991 | 75.0 | 70.4 | 75.0 | 69.9 | 67.6 | 0.473 | 0.505 | 0.473 | 0.514 |
| 10 | 25 | 0.992 | 0.992 | 0.992 | 0.992 | 93.4 | 90.6 | 93.4 | 90.4 | 87.7 | 0.381 | 0.416 | 0.381 | 0.419 |
| 25 | 3 | 1.048 | 1.048 | 1.048 | 1.062 | 78.8 | 78.1 | 78.8 | 77.0 | 39.5 | 0.439 | 0.443 | 0.439 | 0.452 |
| 25 | 10 | 0.927 | 0.927 | 0.927 | 0.934 | 98.2 | 98.1 | 98.2 | 97.9 | 84.8 | 0.300 | 0.308 | 0.300 | 0.311 |
| 25 | 25 | 1.013 | 1.014 | 1.013 | 1.014 | 99.9 | 99.9 | 99.9 | 99.9 | 99.0 | 0.253 | 0.261 | 0.253 | 0.261 |
| 50 | 3 | 1.038 | 1.038 | 1.038 | 1.092 | 98.1 | 97.9 | 98.1 | 97.7 | 36.8 | 0.305 | 0.307 | 0.305 | 0.317 |
| 50 | 10 | 0.906 | 0.906 | 0.906 | 0.912 | 100.0 | 100.0 | 100.0 | 100.0 | 92.8 | 0.212 | 0.215 | 0.212 | 0.217 |
| 50 | 25 | 1.097 | 1.097 | 1.097 | 1.097 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 0.181 | 0.184 | 0.181 | 0.184 |

Figure 3 shows that LM non-coverage was close to 10% when $\rho = 0$. It was very high otherwise as shown by Figures 4 and 5. Hence, use of LM without at least checking $H_0 : \sigma_b^2 = 0$ is not a viable strategy.

Figures 6 - 8 show that confidence intervals using the LM strategy are the shortest, however this strategy is not viable because of its high non-coverage when $\rho \neq 0$. The Huber based approach gives the widest confidence intervals in general. The ADM and ADH confidence intervals are almost always similar to the LMM and Huber ones, respectively. When there were 2 PSUs it is very clear that ADM and ADH confidence intervals are much shorter than LMM and Hub confidence intervals, for all values of $\rho$.

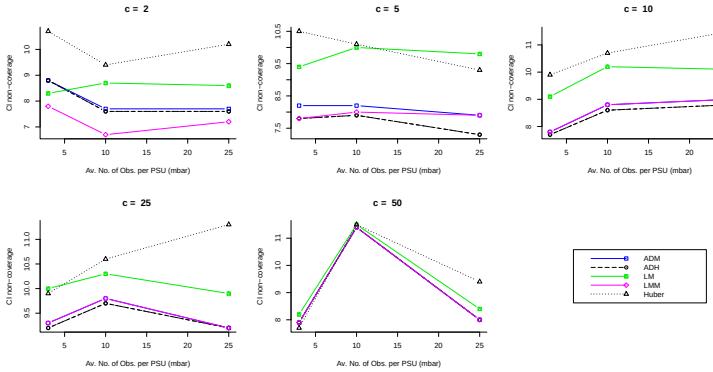



Figure 3: Confidence interval non-coverage rates (%) for different variance estimation methods and various numbers of PSUs ($c$) and units per PSU ($m$) for intraclass correlation ($\rho$) of 0.

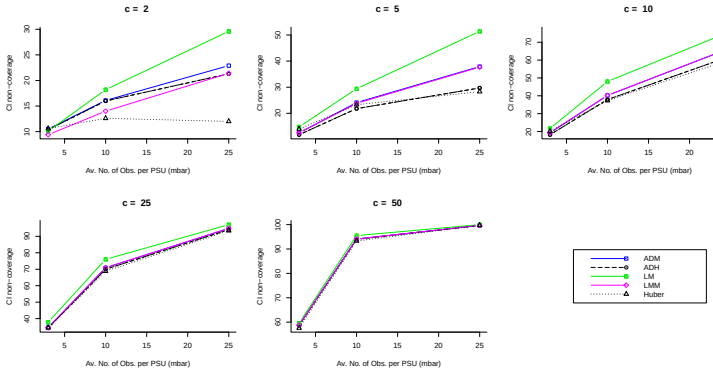



Figure 4: Confidence interval non-coverage rates (%) for different variance estimation methods and various numbers of PSUs ($c$) and units per PSU ($m$) for intraclass correlation ($\rho$) of 0.025.

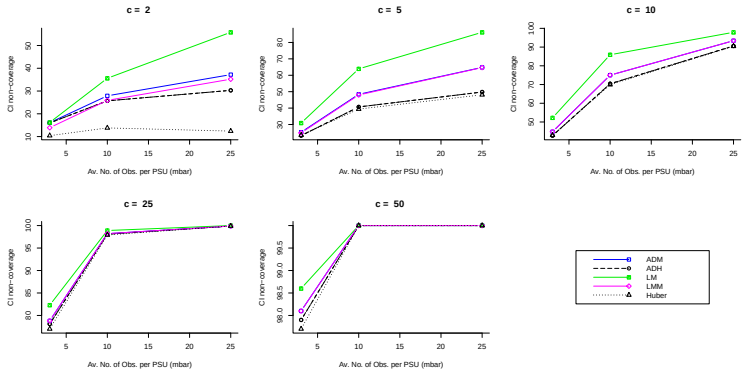



Figure 5: Confidence interval non-coverage rates (%) for different variance estimation methods and various numbers of PSUs ($c$) and units per PSU ($m$) for intraclass correlation ($\rho$) of 0.1.

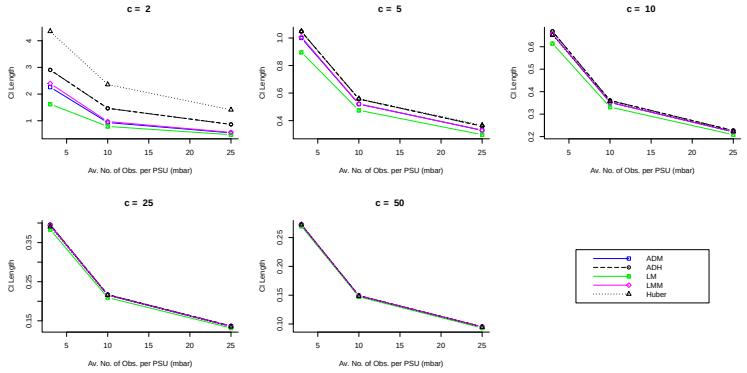



Figure 6: Confidence interval non-coverage rates (%) for different variance estimation methods and various numbers of PSUs ($c$) and units per PSU ($m$) for intraclass correlation ($\rho$) of 0.

## 5 Conclusion

i) Adaptive confidence intervals can perform poorly in designs with few sample PSUs. In these designs, even a small intraclass correlation will substantially inflate the variance of the mean, however the PSU-level variance component is unlikely to be statistically significant even if the intraclass correlation is as high as 0.1. As a result, when the number of PSUs ($c$) is 2 or 5, and the number of observations per PSU ($m$ or $\bar{m}$) is 25 or more both of the adaptive estimators have higher than desirable non-coverage when the intraclass correlation is non-zero, of the order of 15-20%. It appears that for these extreme designs, clustering must be allowed for in variance estimates, even if it is not statistically significant.

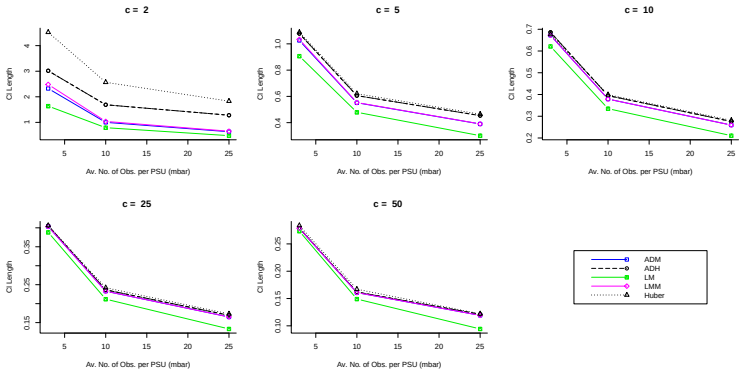



Figure 7: Confidence interval non-coverage rates (%) for different variance estimation methods and various numbers of PSUs ($c$) and units per PSU ($m$) for intraclass correlation ($\rho$) of 0.025.

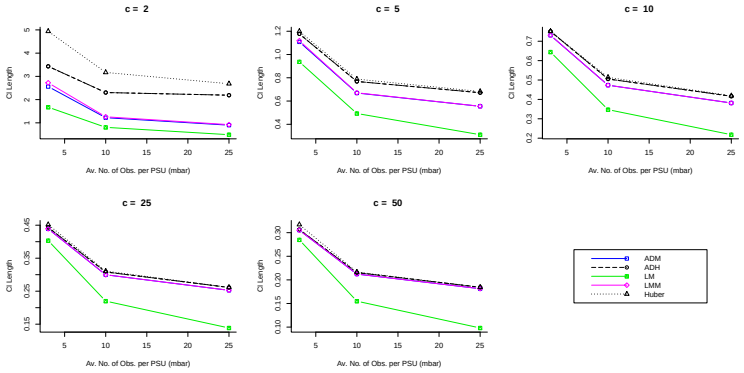



Figure 8: Confidence interval non-coverage rates (%) for different variance estimation methods and various numbers of PSUs ($c$) and units per PSU ($m$) for intraclass correlation ($\rho$) of 0.1.

ii) In comparing the Linear Mixed Model with the adaptive version (ADM), we find that:

- Both the LMM and ADM approaches have close to nominal non-coverage, except for extreme designs of the kind discussed in i. For these designs, the adaptive and non-adaptive LMM methods both have high non-coverage. In the case of the adaptive method, this is presumably because there is not much power to detect the PSU-level variance component, even when it is substantial. For the non-adaptive LMM, the problem seems to be that the LMM confidence intervals are not exact and do not do well for small sample sizes.

- The ADM confidence intervals are noticeably narrower (10-20%) than the LMM for $c = 2$ and 5, but there is not much to choose between ADM and LMM for $c$=10 or more.

iii) In comparing the robust Huber-White approach with the adaptive version (ADH), we find that:

- The Huber approach has close to nominal non-coverage in all cases. So does the ADH approach, except for the extreme designs mentioned in i.
- The Huber method gives wide confidence intervals when $c$ is small (2 or 5) with order of 10-80% eventhough the non-coverage is close to the nominal 10%. This is because the degrees of freedom for this method is equal to ($c$-1). ADH has much narrower confidence intervals (10-80%) , because its degrees of freedom are equal to ($n$-1) rather than ($c$-1) if the PSU-level variance component is not significant.

**Recommendations**

Designs with fewer than 10 PSUs, and a large sample size in each PSU should be avoided, even if the intraclass correlation is believed to be low. Hence, we recommend ignoring clustering if the PSU-level variance effect is insignificant.

## References

Chernoff, H. (1954). On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics*, 25(3):573–578.

Clark, R. G. and Steel, D. G. (2002). The effect of using household as a sampling unit. *International Statistical Review*, 70(2):289–314.

Faes, C., Molenberghs, H., Aerts, M., Verbeke, G., and Kenward, M. G. (2009). The effective sample size and an alternative small-sample degrees-of-freedom method. *The American Statistician*, 63(4):389–399.

Goldstein, H. (2003). *Multilevel Statistical Models*. Kendall's Library of Statistics 3. Arnold, London, third edition.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under non-standard conditions. *Proceedings of the Fifth Berekley Symposium on Mathematical Statistics and Probability, University of California, Berekley*, 11:221–233.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Rao, P. S. R. S. (1997). *Variance Components Estimation, Mixed Models, Methodologies and Applications*. Chapman and Hall.

Scott, A. J. and Holt, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *journal of the American Statistical Association*, 77(380):848–854.

Self, S. G. and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood

estimators and likelihood ratio tests under nonstandard conditions. *journal of the American Statistical Association*, 82(398):605–610.

Stram, D. O. and Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50(4):1171–1177.

Verma, V., Scott, C., and O'Muircheartaigh, C. (1980). Sample design and sampling errors for the world fertility survey. *journal of the Royal Statistical Society, Series A (General)*, 143(4):431–473.

West, B. T., Welch, K. B., and Galecki, A. T. (2007). *Linear Mixed Model: A Practical Guide Using Statistical Software*. Chapman and Hall/CRC, Boca Raton, Florida.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25.